PROBABILISTIC TEMPORAL-LABEL AGGREGATION FOR UNINTENTIONAL ACTION LOCALIZATION

Nuoxing Zhou[†], *Guangyi Chen*[†], *Jinglin Xu*[†], *Wei-Shi Zheng*[‡], *and Jiwen Lu*[†,*]

[†]Department of Automation, Tsinghua University
[‡]School of Computer Science and Engineering, Sun Yat-Sen University

ABSTRACT

Humans can easily understand whether a person's action is intentional or not. However, it is very challenging to teach a machine to recognize this due to the lack of referable comparisons and reliable annotations. Given a video with unintentional action, the annotations are usually unreliable due to the intrinsic ambiguity from multiple annotators and the subjective appraisals. To address this problem, we propose a new framework which online aggregates multiple probabilistic labels for unintentional action localization. Specifically, we first model the uncertainty of annotations with a temporal probability distribution, and then develop a label attention model to aggregate the reliable annotations in an online manner. We evaluate our method on the public OOPS dataset where each video contains multiple annotations of unintentional action and our experimental results show that mining reliable supervision information from multiple unreliable annotations achieves significant improvements over the baseline methods.

Index Terms— Action localization, Unintentional action, Probabilistic label aggregation

1. INTRODUCTION

> *"Intention is one of the most powerful forces there is. What you mean when you do a thing will always determine the outcome."*
>
> *–Brenna Yovanoff, The Replacement*

Existing human action analysis and recognition systems tell us what the contents of physical motions are (action recognition [1]) or when the actions begin and end (action detection [2] and action localization [3]), which cannot explain why the action fails. Hence, it is desirable to require the model understand the intention behind the observed actions, such as raising a glass of wine to one's lips is intentional to drink while spilling the wine all over one's shirt is unintentional action.

The research shows that 18-months-old children are capable of understanding intentional acts [4]. However, it is still an enormous challenge to teach the model to understand the intention of observed actions, which requires referable comparisons and reliable annotations. To train the model for recognizing intention, Epstein *et al.* [5] collect an annotated video dataset with unintentional action, which annotates videos with the temporal location at which the video transitions from intentional to unintentional action. One can train the model to localize the unintentional action by classifying the given video clips (or frames) are intentional or not. Despite there are abundant annotations in the datasets, the an-

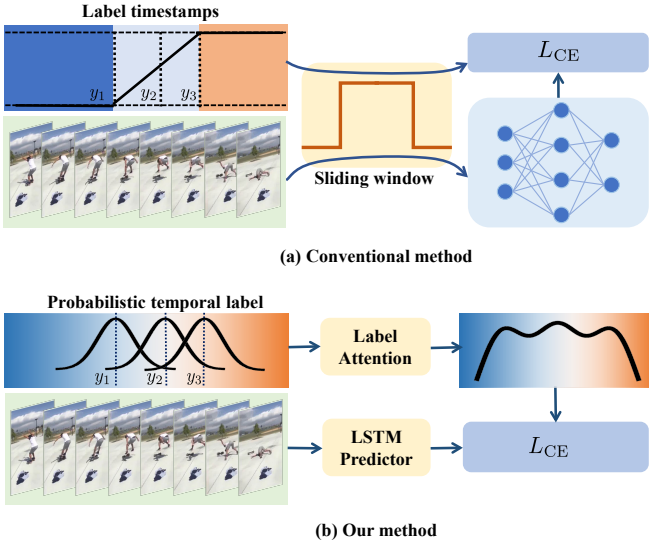

Fig. 1: The comparisons between conventional methods and our proposed method. The top row shows that conventional methods directly use the hard labels to segment the video with three parts including intentional action, transitions from intentional to unintentional action, and unintentional action. This supervision manner ignores the uncertainty of annotations and is easy to be misled by the noisy annotations. The bottom row shows that our proposed method models the label uncertainty with a probability distribution and mines the reliable annotations from multiple candidates by using the label attention model. Through relaxing the hard label and exploring the reliable annotation, our method shows a more powerful generalization ability.

∗Corresponding author

notations are always unreliable due to the intrinsic ambiguity from multiple annotators and their subjective appraisals. We show this intrinsic ambiguity in Fig 3, e.g. for the same video, "a man falls off when he jumps onto the bed", different annotators have different arguments such as the beginning of the jumping or the moment of the man touching the bed. As shown in Fig. 1, exiting methods train the model with 3-way classification (intentional, transitional, and unintentional), which regards the action as the transitional one if the video clip overlaps with any annotated point. It is a hard classification for intentional/unintentional actions, which applies the upper and lower bounds of annotations as the final supervision. The model is easy be misled by the noise of annotations due to over-fitting with this hard supervision.

In this paper, we formulate the unintentional action localization as a temporal probabilistic regression problem, and propose to online aggregate multiple annotations using an attention model. As shown in Fig. 1, we directly regress the timestamp to localize the unintentional action. To model the uncertainty of annotations, we apply a probability distribution, i.e. a Gaussian distribution or a Laplace distribution, to replace the fixed temporal location. In this temporal label distribution, unintentional actions more likely occur in the locations closer to the annotated timestamps. In addition, we propose a label attention model to aggregate the labels from different annotators. This label attention model estimates the reliabilities of different labels and reweights them before probabilistic superposition. Finally, we normalize the aggregated distribution as the supervision to train the model. The generated label distribution considers the uncertainty of annotations, constructs the graduality from the intentional action to unintentional action, and mines the reliable clues from the multiple unreliable annotations. We evaluate our method on the OOPS [5] dataset and obtain significant improvement.

2. RELATED WORK

In this section, we briefly review two related topics: unintentional action localization and label distribution learning.

2.1. Unintentional Action Localization

Different from conventional action localization [6, 7, 8, 9] which focuses on the beginning and ending of the action contents of the video, unintentional action localization aims at understanding the intention behind the action and localizing when the action becomes unintentional. To understand the intention, Epstein *et al.* [5] collect an annotated video dataset and train a three-way classifier to recognize the action as intentional, unintentional, or transitional. It localizes the unintentional action by applying the classifier in a sliding window fashion over the temporal axis and exploring the location with the most confident score. Furthermore, the goals of original intentional action are labeled to improve the quality of the supervision and train the more discriminative video representations [10]. However, these models are easy to be misled by unreliable annotations due to the hard supervision manner. Thus, we propose to aggregate label distributions from multiple annotations online.

2.2. Label Distribution Learning

Label distribution learning [11, 12, 13] aims to solve the uncertainty of annotations by replacing a hard label with a probability distribution, which has obtained great success for facial age estimation. For example, Geng *et al.* [11] first propose to apply an age distribution as the supervision instead of a fixed age label, and extend it into deep learning framework [12]. Recently, label distribution learning has widely used in different computer vision tasks such as facial landmark detection [14], pose estimation [15] and crowd counting [16], and demonstrates the effectiveness by mitigating the overfitting of unreliable annotations. In this work, we apply the label distribution for the temporal location of the video, and further, propose an attention model to online aggregate multiple label distributions from different annotators.

3. APPROACH

This section presents a temporal probabilistic regression framework to study unintentional action localization, where the core idea consists of probabilistically modeling unreliable temporal annotations and online aggregating multiple labels via attention model.

3.1. Problem Formulation

The unintentional action localization task aims to detect the temporal boundary between intentional and unintentional action. Given a video including T frames $X = \{x_t|t = 1, 2, ..., T\}$, we formulate the unintentional action localization task as the temporal regression in which the model predicts the temporal location when unintentional action occurs. To obtain this location, we predict the probabilities of unintentional action occurring for each frame in the video and select the most likely time as the predicted temporal location:

$$\hat{y} = \max_t q_\theta(x_t), \tag{1}$$

where q_θ denotes the prediction model whose input is the representations of frames (or video clips) and output is the predicted probability of unintentional action occurring. However, the annotations of unintentional action are always unreliable due to the intrinsic ambiguity from multiple annotators and their subjective appraisals, i.e. we have different labeled failure moments $Y = \{y_k|k = 1, 2, ..., K\}$ from different annotators for the same video in the OOPS dataset. Thus, a challenging problem for training the model is how to generate reliable supervision from these unreliable annotations.

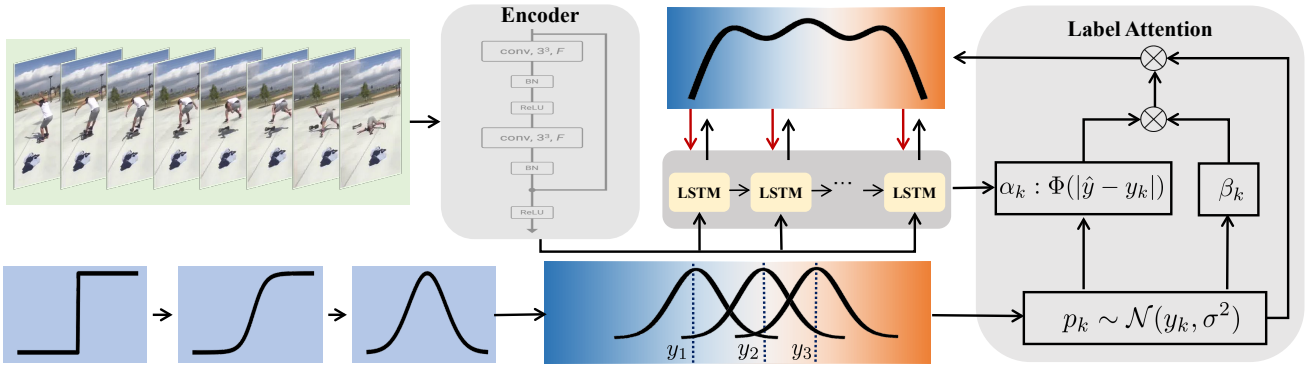

Fig. 2: The overall framework of the proposed method. Given a video with multiple annotations, the model first extract features with a video encoder and apply an LSTM predictor to localize the unintentional action. To model the uncertainty of annotations, we use a label distribution to replace the hard label, which can be regarded as the derivative of the sigmoid approximation of the original signum function. During training, we apply a label attention model to online aggregate multiple annotations by the comparisons between model prediction and annotations and encourage the annotations close to the prediction.

In this work, we propose to soften the hard label information along the temporal axis and further take the derivative of the soft label information to model each failure moment as a probabilistic temporal label distribution. Based on various generative label distributions, we propose to aggregate them in an online manner, which enables our model to focus on more reliable annotations via the attention model.

3.2. Online Probabilistic Temporal Label Aggregation

Given any labeled failure timestamp $y_k \in Y$, we first soften this hard label into a Gaussian distribution as :

$$p_k(t) \sim \gamma \mathcal{N}(\mu = y_k, \sigma^2) \tag{2}$$

where t denotes different temporal locations, the original labeled timestamp y_k is the mean of the distribution, variation σ denotes the degree of deviation, $\gamma = \sqrt{2\pi}\sigma$ is a normalization term which adjusts the $p_k(y_k) = 1$. We use $p_k(t)$ to represent the probability that the unintentional action occurs on this temporal location t. As shown in Fig. 2, this label distribution is equivalent to a two-stage refinement which first approximates the hard signum function with a sigmoid function and then calculates its derivative as the label distribution, since the original symbolic function is non-differentiable. The $p_k(t)$ is larger when the temporal location is close to the labeled timestamp y_k. Besides, we only consider the $p_k(t)$ in the domain of definition $[0, T]$. Note that, we can replace the Gaussian distribution with any unimodal and symmetric distribution, e.g. Laplace distribution.

Despite we model the uncertainty with the probability distribution, how to mine the reliable supervision from multiple annotations is still a challenge. To explore which annotation is more reliable, we propose an attention model to online aggregate multiple annotations. Given a video X and its corresponding annotations Y, the model first predicts the location $\hat{y}$ of unintentional action as (1), and then learns the distance between the prediction and annotations as:

$$\alpha_k = \Phi(|\hat{y} - y_k|), \tag{3}$$

where Φ is a negative correlation function to explore reliable annotations and reweight them. With this simple yet effective label attention model, we pay more attention to the annotations which are close to the model prediction. We online aggregate annotations by this attention model which mines the reliable annotations with the model to learn. Suppose we have an initial weight β_k for each annotation, we will aggregate the probabilistic temporal labels as the final training label:

$$p_y = \sum_{k=1}^{K} \frac{\alpha_k \beta_k p_k}{\sum_{k=1}^{K} \alpha_k \beta_k}, \tag{4}$$

where $\frac{\alpha_k \beta_k}{\sum_{k=1}^{K} \alpha_k \beta_k}$ can be regarded as the posterior weight which is modified by the model prediction. Finally, we also normalize the $p_y \in [0, 1]$.

During the training process, we apply the softmax function to calculate the predicted probability and optimize the model with a cross-entropy loss between predicted probability by the network and generated ground-truth probability by the attention model:

$$\mathcal{L}_{CE}(p_y || q_\theta) = \sum_{t=1}^{T} p_y(t) \log q_\theta(x_t). \tag{5}$$

3.3. Network

In this subsection, we introduce the network architecture of our method, which consists of three parts: video encoder, LSTM predictor, and label attention model. As shown in Fig. 2, for a given video, we first extract features of the frames (or video clips) with an encoder, i.e. an R3D network. Then we

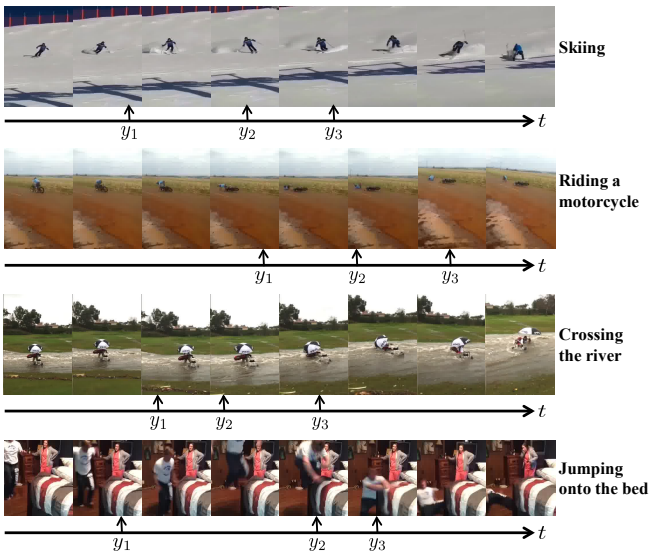

Fig. 3: Examples of the OOPS dataset. We show some example videos of unintentional actions in the OOPS dataset including falling down while skiing, falling down while riding a motorcycle, washed away when crossing the river, and falling off when jumping onto the bed. We also show the annotated temporal locations of the transition from intentional to unintentional action, which is intrinsic ambiguous due to the subjective appraisals of different annotators.

learn an LSTM predictor whose input is the video representations to predict the frame-level probability of unintentional action occurring in the corresponding timestamp. Besides, the label attention model consists of two inputs including multiple probabilistic annotations and model predictions. The label attention model is trained to mine the reliable annotations and aggregate multiple annotations with learned reliability scores. This aggregation process is online updated with different model predictions as inputs.

4. EXPERIMENTS

In this section, we evaluated our proposed method on the OOPS dataset to demonstrate the effectiveness of our proposed method for localizing unintentional action. We also conducted some quantitative comparisons with other methods and analyzed the compared results qualitatively.

4.1. Dataset

The OOPS dataset contains unintentional actions caused by various errors and factors, which is a big amount of collection of videos consisting of over 20000 videos from the YouTube website. In this dataset, there are 4673 labeled training videos and 3593 testing videos, where each labeled video is annotated with the time-stamp at which the video frame begins to happen the unintentional action. Furthermore, according to the statistical information of the OOPS dataset, fifty percent of videos are mainly between the five-second and ten-second, and forty percent of videos start the key unintentional actions in the middle length of the video. The mean video clip length is 9.4 seconds. We show some examples of the OOPS dataset in Fig. 3, such as a man falls off when he jumps onto the bed. We can observe that annotations of when the intentional action transitions to the unintentional action are intrinsic ambiguous. For example, one annotator argues the unintentional action occurs at the beginning of the jumping, while other annotators argue it occurs when the man touches the bed.

Table 1: Comparisons of unintentional action localization of baseline and our method on the OOPS dataset.

Method	Localization Accuracy	
	within 1 sec	within 0.25 sec
PUAV-Chance [5]	25.9	6.8
PUAV-VideoSort [5]	43.3	18.3
PUAV-VideoContext [5]	52.0	25.3
PUAV-VideoSpeed [5]	65.3	36.6
PUAV-Pretrain [5]	69.2	37.8
GEWF [10]	72.4	39.9
Ours-Hard	69.8	38.0
Ours-Probabilistic	71.6	38.4
Ours-Online	**73.2**	**40.2**

4.2. Experiment Settings

During the training phase, we clipped each video to 90 video frames, utilized the 3DResNet-18 [17] pretrained on Kinetics [18] to extract 512-dimension visual features for each video frame at the last convolutional layer. After that, we applied a 2-layer basic LSTM as the backbone and used Adagrad optimizer to train the model with an initial learning rate of 0.001, where the dimensions of input, hidden state, and output are 512, 128, and 2, respectively. Note that, we did not finetune the backbone network for a fair comparison.

During the test phase, we directly extracted 512-dimension visual features for each untrimmed video and fed them into our learned model to predict all the video frame labels which are utilized to localize the timestamp of happening unintentional action. Based on the predicted temporal label distribution of the video frames, we followed the evaluation setting of localization in [5]. Specifically, we used our model in a sliding window fashion over the temporal axis and evaluated whether the model can detect the timestamp of the unintentional action beginning. The predicted boundary is the one with the most confident score of unintentional action across all the sliding windows. We considered the prediction correct if the predicted boundary sufficiently overlaps any of the ground truth positions in the dataset, where two different thresholds of sufficient overlap are utilized, i.e., within

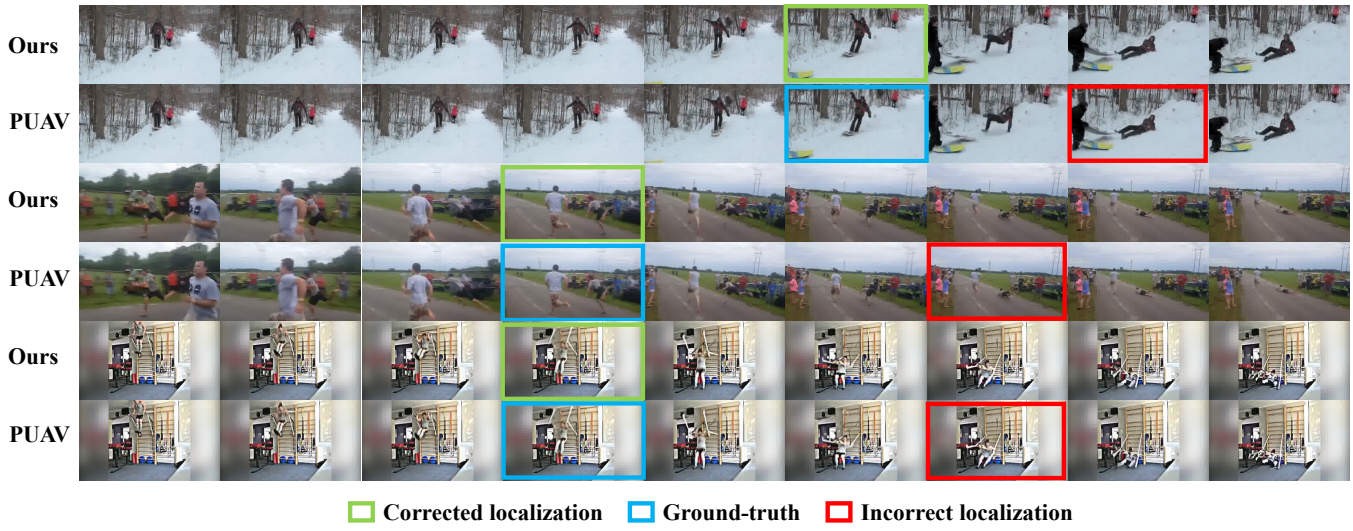

Fig. 4: The result comparisons between our method and PUAV [5].

one-second and within the one-quarter-second. Note that we did not follow the classification settings in [5], since we only regressed the probability of unintentional action occurring instead of the three-way classification. Our implementation was based on PyTorch and our hardware configuration comprised a 3.70GHz CPU and 31GB RAM. We used an NVIDIA 2080Ti GPU for neural network acceleration.

4.3. Quantitative Analysis

We compared our method with the methods used in [5], including VideoSpeed, VideoContext, VideoSort, and the predicted model on Kinetics, and the method GEWF [10] which uses the extra annotations of action goals. These compared methods utilized the model pre-trained on the full, annotated Kinetics [18] dataset as feature extractors.

Table 1 shows the results of the task of unintentional event localization. It is observed that our approach outperforms other compared methods. For example, making a comparison between Ours-Online and the PUAV-Pretrain methods with the same pretrained features, we can obtain 4.2 percent and 2.4 percent improvements on both settings of within one second and within a one-quarter second, which indicates that our label attention model is more appropriate to localize the unintentional actions by using fine-grained frame-level prediction. Besides, Ours-Online also outperforms the GEWF method on both settings of within one second and within a one-quarter second, even though we did not use the extra annotations of the goal to improve the quality of supervision. It demonstrates that constructing the temporal probabilistic regression framework and label attention model has the ability to capture more informative perceptual clues to localize unintentional actions.

In addition, we investigated the critical factors beneficial to localizing unintentional actions by ablation studies. We evaluated different versions of our method in both settings, i.e., Ours-Hard, Ours-Probabilistic, and Ours-Online methods, respectively, within one second and within a one-quarter second, which analyzes the effects of different label models. Specifically, the Ours-Hard method degraded the probabilistic temporal label to the simple scalar labels without considering the label attention model. The Ours-Probabilistic method used a probabilistic temporal label to supervise the learning model without using the attention model to aggregate online. Compared with the PUAV-Pretrain method, the Ours-Hard model used the same pretrained features replacing the linear classification with an LSTM predictor. Comparing the performance of the PUAV-Pretrain method and the Ours-Hard model show that the improvement of modifying the model architecture is trivial. While making comparisons between the Ours-Hard model and the Ours-Probabilistic method show that modeling the uncertainty of annotations as the temporal label distribution is effective. Furthermore, we evaluated the performance of the proposed attention-based online aggregation method by comparing the Ours-Probabilistic method and the Ours-Online method. The further improvement demonstrates that our method has the ability to mine reliable supervision information from multiple annotations.

4.4. Qualitative Analysis

To investigate the effectiveness of our proposed method, we also conducted the qualitative analysis by visualizing the localization results of PUAV [5] and our method. We showed the comparisons and fail examples respectively in Fig. 4 and Fig. 5. Taking the first video "falling down while skiing" as an example for comparison, our method correctly localized the unintentional action while PUAV failed on the "fallen man". It indicated that the model trained by the hard labels always overfit to the fallen results instead of the unintentional causes like "sloping skis". Besides, we also showed some failed

Fig. 5: The fail examples of our method.

examples of our method. For example, the model still tends to capture the large visual changes of video (e.g., "the man falls off a cow" in the first video), but not the real beginning of unintentional action ("the cow is crazy"), despite mining reliable supervision. It encourages us to further explore the causalities of unintentional action.

5. CONCLUSION

In this paper, we have proposed a probabilistic temporal-label aggregation method for unintentional action localization, which replaces the hard category labels with a temporal probability distribution and online aggregates multiple annotations through an attention model. We formulate the uncertainty of annotations as a prior distribution and learn the label attention model to estimate the reliabilities of multiple labels and accordingly reweight them. We have demonstrated significant improvements over baseline methods with the proposed method.

6. ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant U1813218, Grant 61822603, Grant U1713214, in part by Beijing Academy of Artificial Intelligence (BAAI), in part by a grant from the Institute for Guo Qiang, Tsinghua University, and in part by China Postdoctoral Science Foundation under Grant 2020M680564.

7. REFERENCES

[1] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NeurIPS*, 2014, pp. 568–576.

[2] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei, "End-to-end learning of action detection from frame glimpses in videos," in *CVPR*, 2016, pp. 2678–2687.

[3] Zheng Shou, Dongang Wang, and Shih-Fu Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *CVPR*, 2016, pp. 1049–1058.

[4] Andrew N Meltzoff and Rechele Brooks, "Like me" as a building block for understanding other minds: Bodily acts, attention, and intention," *Intentions and intentionality: Foundations of social cognition*, vol. 171191, 2001.

[5] Dave Epstein, Boyuan Chen, and Carl Vondrick, "Oops! predicting unintentional action in video," in *CVPR*, 2020, pp. 919–929.

[6] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar, "Rethinking the faster r-cnn architecture for temporal action localization," in *CVPR*, 2018, pp. 1130–1139.

[7] Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S Davis, and Yan Qiu Chen, "Temporal context network for activity localization in videos," in *ICCV*, 2017, pp. 5793–5802.

[8] Huijuan Xu, Abir Das, and Kate Saenko, "R-c3d: Region convolutional 3d network for temporal activity detection," in *ICCV*, 2017, pp. 5783–5792.

[9] Jiyang Gao, Zhenheng Yang, Kan Chen, Chen Sun, and Ram Nevatia, "Turn tap: Temporal unit regression network for temporal action proposals," in *ICCV*, 2017, pp. 3628–3636.

[10] Dave Epstein and Carl Vondrick, "Video representations of goals emerge from watching failure," *arXiv preprint arXiv:2006.15657*, 2020.

[11] Xin Geng, Chao Yin, and Zhi-Hua Zhou, "Facial age estimation by learning from label distributions," *TPAMI*, vol. 35, no. 10, pp. 2401–2412, 2013.

[12] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng, "Deep label distribution learning with label ambiguity," *TIP*, vol. 26, no. 6, pp. 2825–2838, 2017.

[13] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *NeurIPS*, 2019, pp. 1567–1578.

[14] Kai Su and Xin Geng, "Soft facial landmark detection by label distribution learning," in *AAAI*, 2019, vol. 33, pp. 5008–5015.

[15] Xin Geng and Yu Xia, "Head pose estimation based

on multivariate label distribution," in *CVPR*, 2014, pp. 1837–1842.

[16] Miaogen Ling and Xin Geng, "Indoor crowd counting by mixture of gaussians label distribution learning," *TIP*, vol. 28, no. 11, pp. 5691–5701, 2019.

[17] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?," in *CVPR*, 2018, pp. 6546–6555.

[18] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017, pp. 6299–6308.